APPLICATION OF MACHINE LEARNING METHODS FOR ANALYZING DATA FROM THE NOMENCLATURE DIRECTORY OF THE ENTERPRISE RESOURCE PLANNING SYSTEM, PART 2

Mushtak O.I.[1], Limanovskaya O.V. [1], Lebedev A.S.[1,2,3*]

[1)] Ural Federal University, Yekaterinburg, Russia
[2)] Ural Mining and Metallurgical Company (UMMC), Yekaterinburg, Russia
[3)] None-state Higher Educational Establishment "UMMC Technical University", Yekaterinburg, Russia
*E-mail: aslebedev@urfu.ru

At the first stage, all units of measurement of materials in the database were reduced to 4 categories: pieces, kg, km and the rest.

Further, in order to avoid duplication of names having the same value, a morphological analysis of the names was performed using pymorphy2. As a result, the names of the materials were cleared of punctuation and brought into normal form.

The resulting text data was translated into a numerical feature vector using the CountVectorizer. Since the names in the overwhelming majority of cases are represented in one word, N-grams were not used.

As a result, the data were presented to the analysis in the form of numerical vectors, and the target variable was the category of units.

Before the analysis selection was divided into training and test (in the ratio 0.8 to 0.2) for further measurement of indicators of classification quality.

The following studies were conducted to identify the optimal parameters of the classifiers.

The Random Forest classifier was tested the effect of the number of estimators on the quality of the model. In the naive Bayes method, the influence of the coefficient of smoothing (additive smoothing) on the quality of the model is investigated. In the xgboosting method, the effect of the depth of learning on the quality of the model is investigated.

The results are shown in Fig.1.

As can be seen from Fig. 1, the quality of the model on the test part of the sample is slightly less than on the train part, which indicates the absence of retraining of the model. At the same time, the quality of the model on the test part of the sample is not very different from that on the train part and remains fairly high, which indicates that the model is well-trained. The increase in the number of trees does not greatly change the quality of the model, there is a slight increase in it, and the maximum value of 0.876 is reached with 25 estimators. A further increase in the number of trees does not increase the accuracy of the model. Therefore, 25 trees are optimal for the Random Forest classifier.

As can be seen from Figure 1, the quality of the model on both the train and test part of the sample is high and no retraining is observed. The maximum quality of the model (accuracy 0.876) is achieved with an alpha value of 0.56.

As can be seen from Fig. 1, the quality of the model both on the test and on the train part of the sample is not very different and is quite high. Re-training of the model is not observed. The quality of the model obtained at xgboost is less than the two previous classifiers. The maximum accuracy of the model (0.863) is achieved with a depth of training equal to 15.

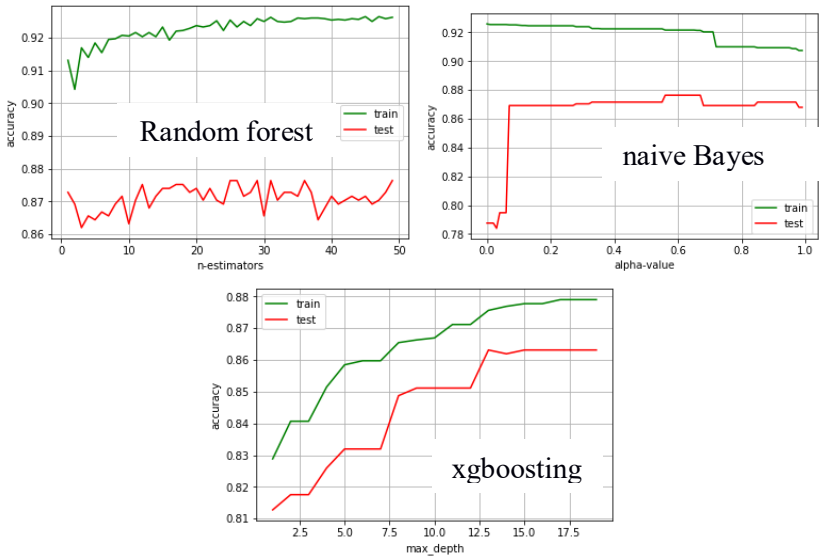

Fig. 1. Results of applying different methods

Result: the random forest and naive Bayes methods are effective for multi-class classifications.

The work done preparing the text data from the export data in CSV-format from corporate directory of materials in the UMMC CIS to add to the analysis and the classification of materials on the basis of units of measurement. Consistently high quality classification method showed a random forest.